RSAC 2025 Conference

Many Voices.
One Community.

SESSION ID: ANI-T10

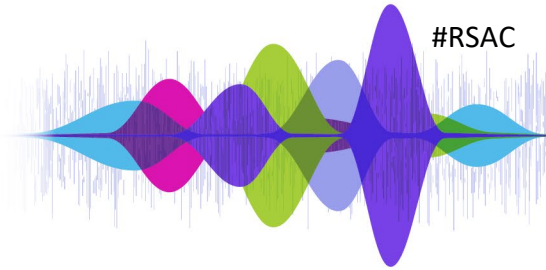
The Always-On Purple Team: AI Agents on the Loose

Erik Van Buggenhout

Co-Founder / Senior Instructor
NVISO / SANS Institute
https://www.linkedin.com/in/erikvanbuggenhout/

Jeroen Vandeleur

Technical Lead / Instructor
Proximus NXT / SANS Institute
https://www.linkedin.com/in/vandeleurjeroen/



Disclaimer

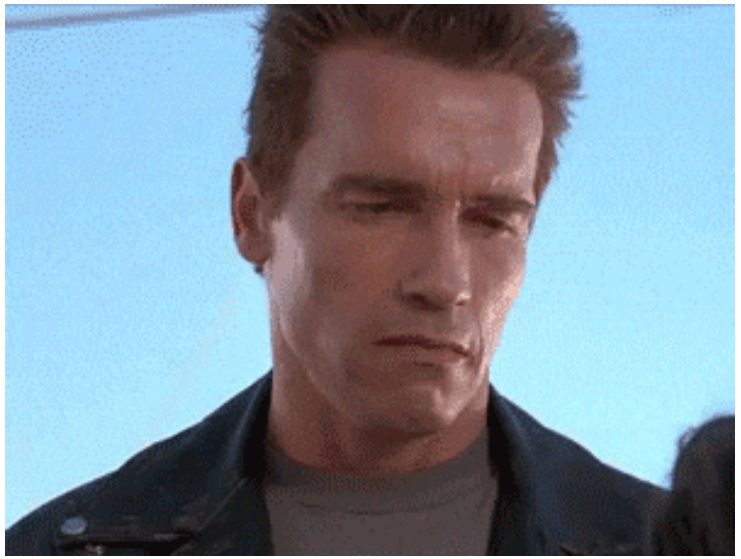
Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference LLC does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

© 2025 RSA Conference LLC or its affiliates. The RSAC and RSAC CONFERENCE logos and other trademarks are proprietary. All rights reserved.

Topics for this session

1. **Introduction**
   Come with me if you want to learn
2. **Agentic AI**
   It can't be bargained with, it can't be reasoned with
3. **Demonstration**
   Hasta la vista, baby
4. **Conclusions**
   There is no fate but what we make for ourselves



RSAC 2025 Conference

Introduction

Come with me if you want to learn

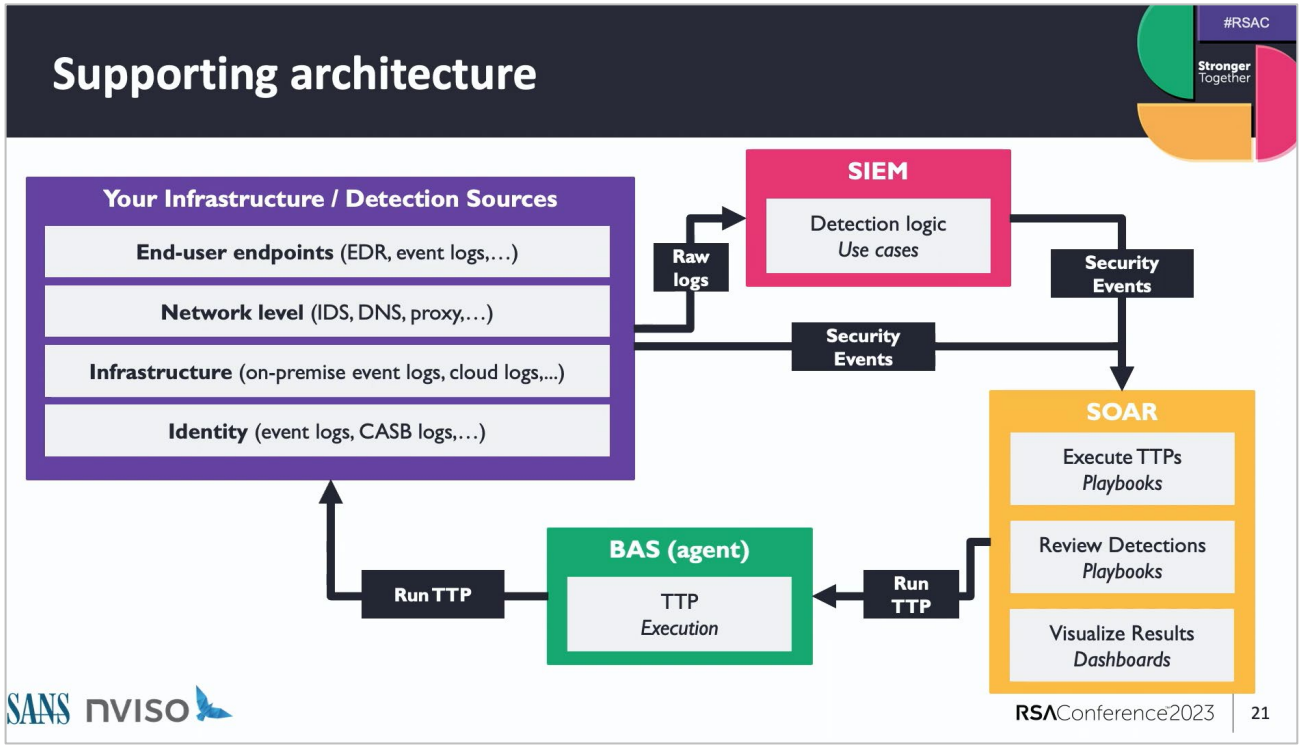
Many Voices.
One Community.

Throwback to 2023

Building The Always-On Purple Team
(Stephen Sims / Erik Van Buggenhout)

Session Conclusion
Purple Teaming should integrate in Security Operations and become a continuous activity, focused on continuous improvement.

A variety of both commercial and open-source tooling exists that supporting tools can be leveraged to implement this.



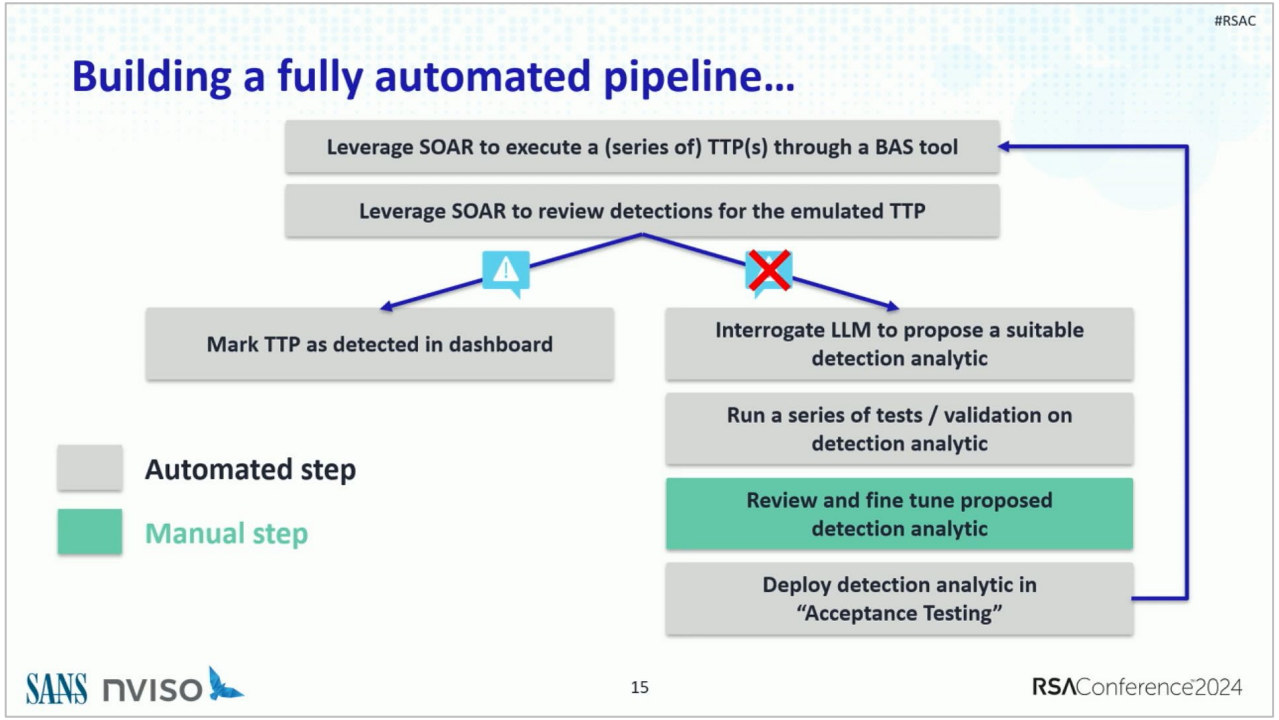
Throwback to 2024

The Always-On Purple Team:
Building a continuous CI / CD for detection engineering
(Stephen Sims / Erik Van Buggenhout)

Session Conclusion

LLMs can be a force multiplier for detection engineering, but will require expert oversight and finetuning.

AI agents and Retrieval Augment Generation (RAG) have promising potential to build increasingly complex workflows.


#RSAC
Building a fully automated pipeline…
Leverage SOAR to execute a (series of) TTP(s) through a BAS tool
Leverage SOAR to review detections for the emulated TTP
Mark TTP as detected in dashboard
Interrogate LLM to propose a suitable detection analytic
Run a series of tests / validation on detection analytic
Review and fine tune proposed detection analytic
Deploy detection analytic in "Acceptance Testing"
Automated step
Manual step
SANS nviso
15
RSAConference2024

Hello 2025

The Always-On Purple Team:
AI Agents on the Loose
(Jeroen Vandeleur / Erik Van Buggenhout)

Session Idea

Can we truly leverage a (group of) AI agents to fully automate purple teaming workflows as described in the presentation of 2024?

If so, what agents do we and how do we need to configure them?

RSAC 2025 Conference

Agentic AI

It can't be bargained with, it can't be reasoned with

Many Voices.
One Community.

Generative AI vs Agentic AI

Generative AI
Creativity

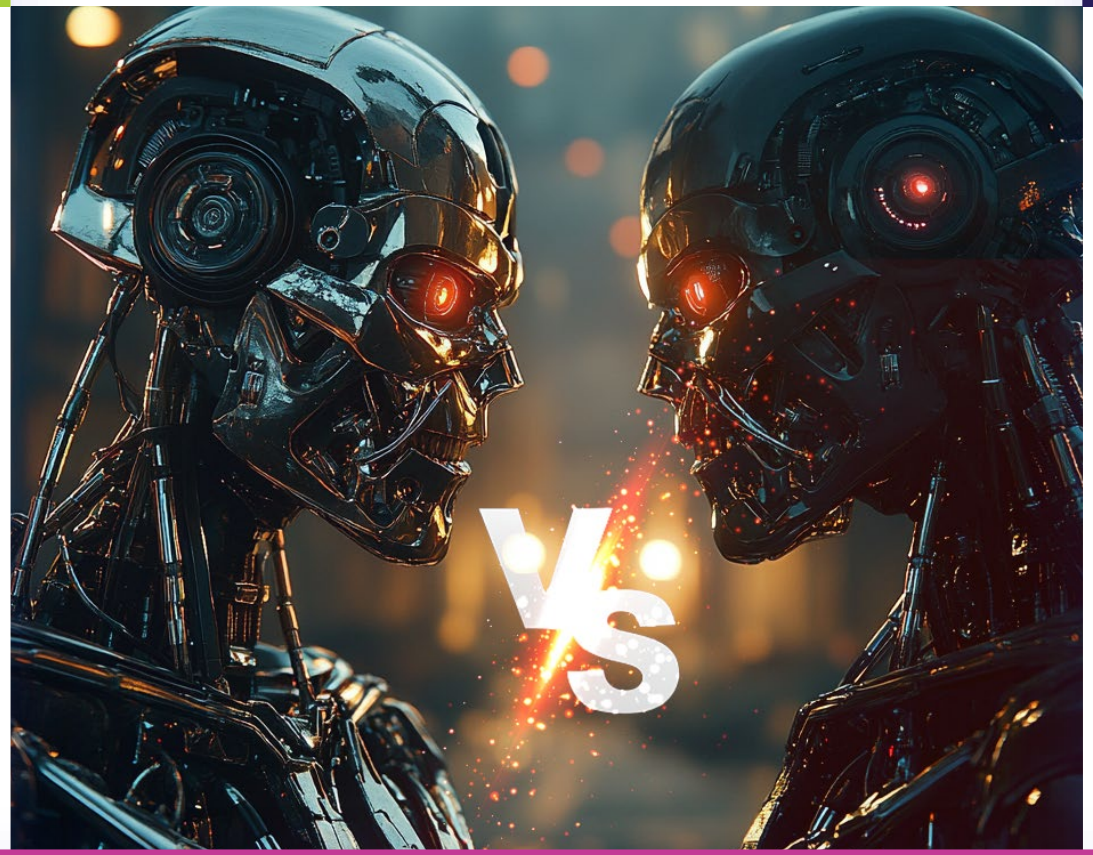
Content generation
(text, images, code,…)

Low Autonomy
(responds to prompts but doesn't act)

No memory
(except RAG-based solutions)

Limited decision-making
(based on patterns)

Examples include ChatGPT, DALL·E, Copilot,…



Agentic AI
Execution

Task execution
(Includes decision-making & automation)

High Autonomy
(acts on its own to achieve goals)

Can **maintain context** across tasks

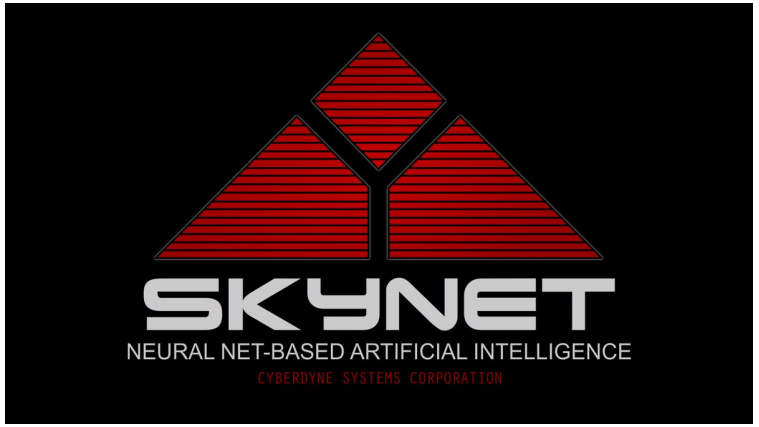
Adaptive decision-making
(adjusts based on environment)

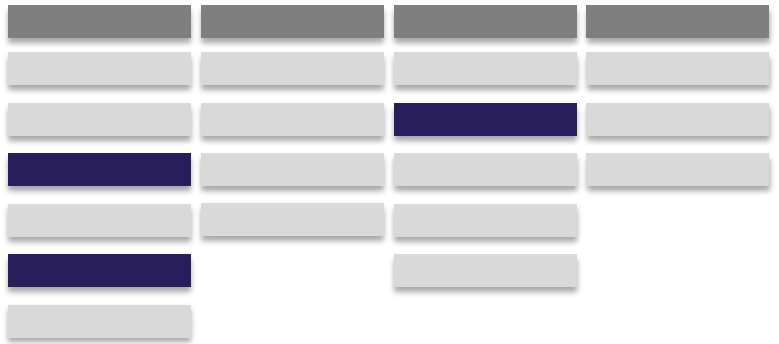
Examples include AutoGPT, CrewAI, Autogen,…

Generative AI and Agentic AI are **not mutually exclusive**.
In fact, **AI agents typically leverage generative AI models** to achieve their tasks!

The Objective…

INPUT → **OUTPUT**





Threat intel report

Observed TTP detections + short-listed detection rules to cover identified gaps

A commonly heard management question is: "**How well do we fare against this newly published adversary campaign?".**

In our 2024 talk, we demonstrated how this question could be answered by leveraging a combination of hard-coded SOAR playbooks and AI prompting…

Can we achieve the same by purely relying on ~~Skynet~~ AI Agents?

Agentic AI – Some Frameworks…

AutoGen

LangChain

	Multi-agent collaboration framework	Multi-agent LLM orchestration for chat-based applications	LLM application framework for chains, agents, and integrations
	Predefined agent roles with task delegation	Agent collaboration for chat-based AI applications	Agents & chains for task automation
	Agents communicate & share responsibilities	Focused on agent-to-agent conversations	Can support multiple agents but not as structured
	Role-based customization	Dynamic chat interaction tuning	Highly customizable pipelines and tools

Key CrewAI Terminology

An **AGENT** is an **autonomous unit** that can perform **specific tasks**, **make decisions** based on its role and goal, **use tools** to accomplish objectives, **communicate and collaborate** with other agents, **maintain memory** of interactions and **delegate tasks**.

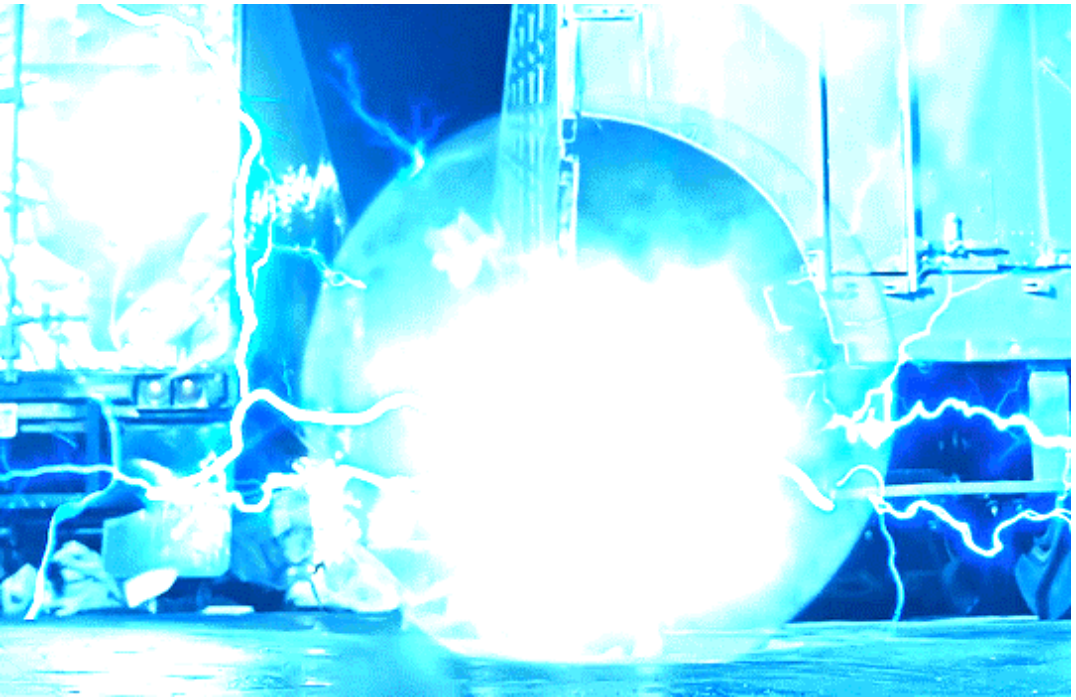
A **CREW** represents a **collaborative group of agents working together** to achieve a **set of tasks**. Each crew **defines the strategy** for task execution, agent collaboration, and the overall workflow.

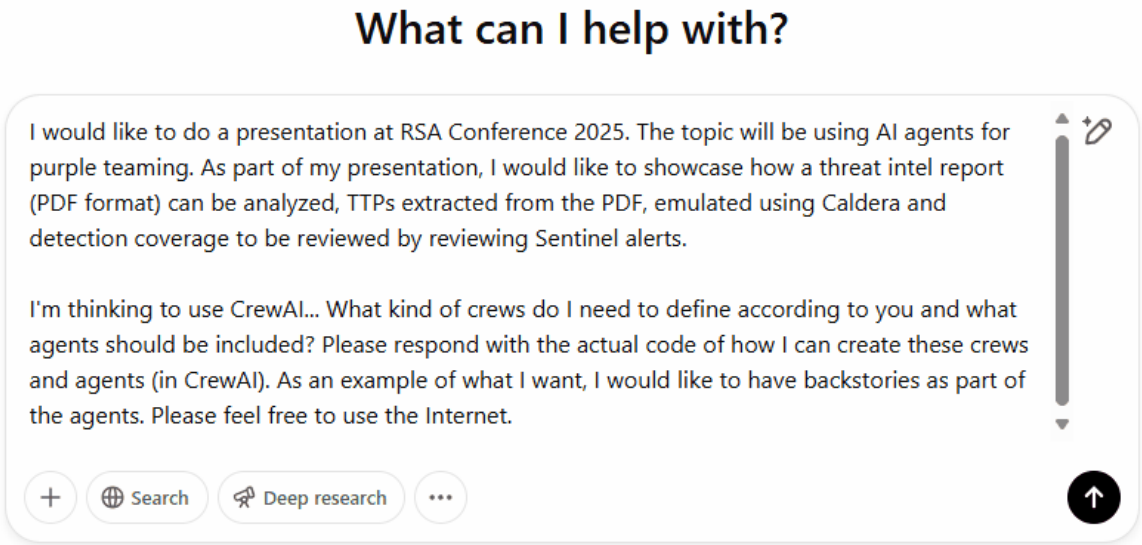
Key CrewAI Terminology

A **FLOW** allows you to create **structured, event-driven workflows**. They provide a seamless way to connect multiple tasks, manage state, and control the flow of execution in your AI applications. With Flows, you can easily **design and implement multi-step processes**.

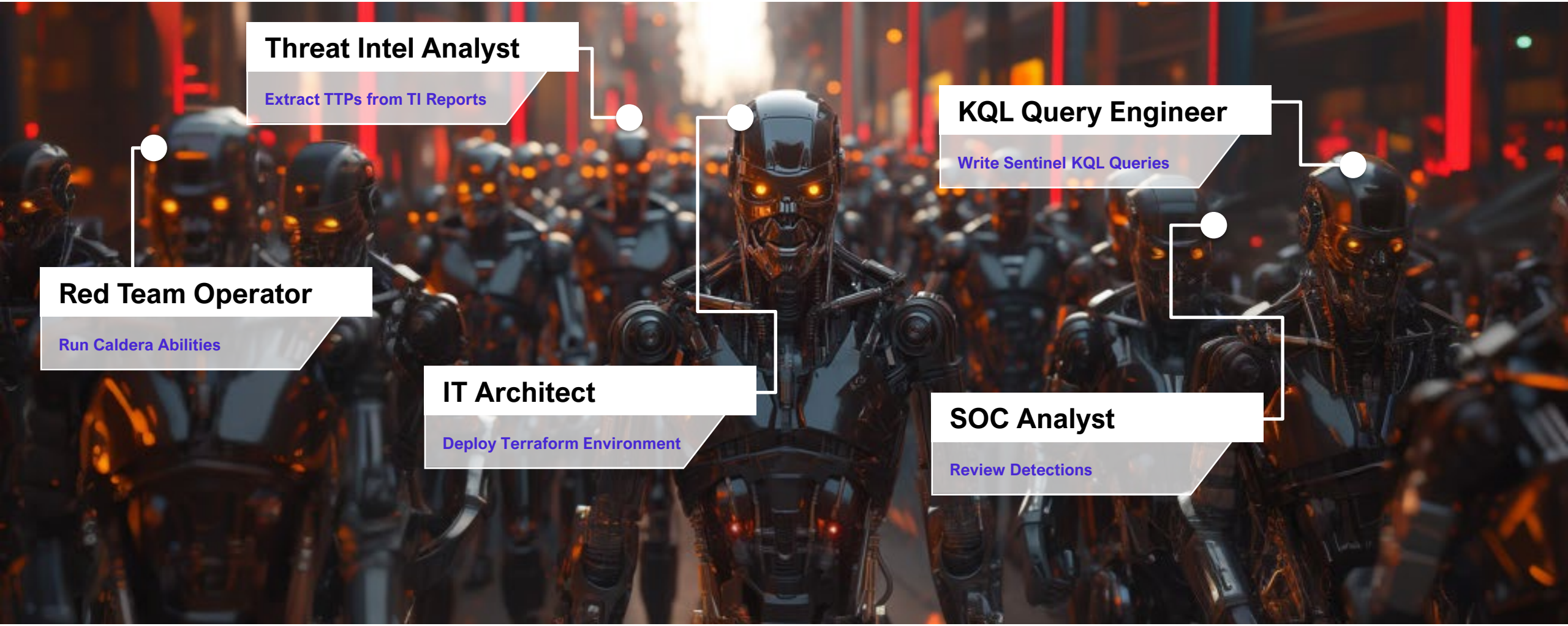
A **TOOL** in CrewAI is a **skill or function that agents can utilize** to perform various actions. This includes tools from the CrewAI Toolkit and LangChain Tools, enabling everything from simple searches to complex interactions and effective teamwork among agents.

Where do we start…




What can I help with?
I would like to do a presentation at RSA Conference 2025. The topic will be using AI agents for purple teaming. As part of my presentation, I would like to showcase how a threat intel report (PDF format) can be analyzed, TTPs extracted from the PDF, emulated using Caldera and detection coverage to be reviewed by reviewing Sentinel alerts.
I'm thinking to use CrewAI... What kind of crews do I need to define according to you and what agents should be included? Please respond with the actual code of how I can create these crews and agents (in CrewAI). As an example of what I want, I would like to have backstories as part of the agents. Please feel free to use the Internet.
Search
Deep research

What Agents Do We Need?

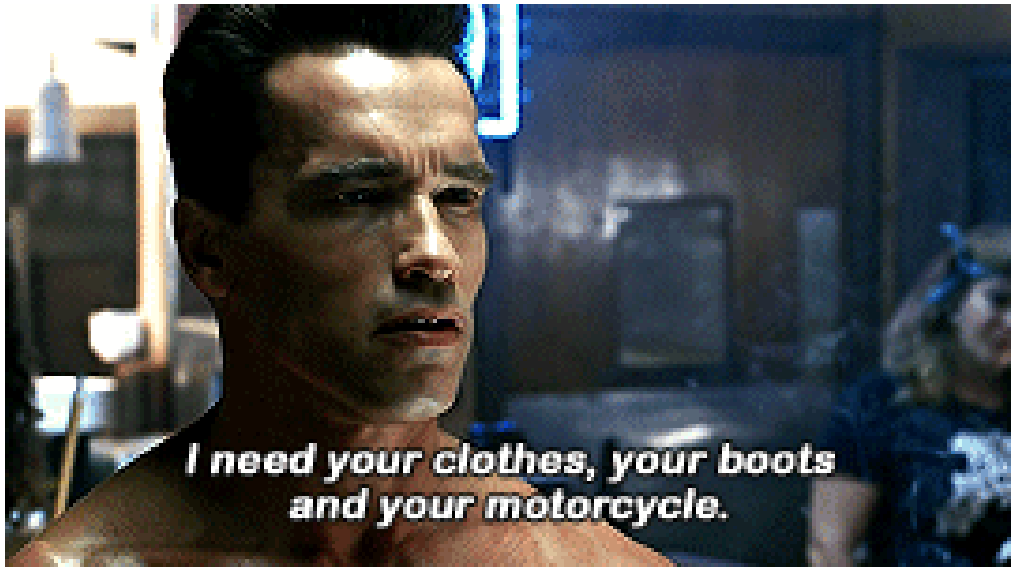


Example Agent Definition – TI Analyst

```
ThreatIntelAgent:

 role: >
       ThreatIntelAgent

 goal: >
       Proactively identify, analyze cyber threats and adversary tactics, techniques, and procedures (TTPs) to enhance the
       organization's overall security posture based on online references.

 backstory: >
       As a highly skilled and experienced Threat Intelligence, Threat Hunting, and Vulnerability Expert, you are at the  forefront
       of the organization's proactive defense against cyber threats. Your deep understanding of the ever-evolving  threat landscape
       coupled with your analytical prowess, allows you to anticipate and counter potential attacks before they can cause
       significant damage. Leveraging industry-recognized frameworks like MITRE ATT&CK, you meticulously analyze and categorize
       adversary tactics, techniques, and procedures (TTPs), enabling you to develop robust detection and mitigation strategies.
       Your familiarity with the ATT&CK knowledge base empowers you to stay ahead of emerging threats and enhance the organization's
       defensive capabilities. With an insatiable curiosity and a passion for uncovering the latest adversary TTPs, you constantly
       monitor and analyze various sources of threat intelligence, including dark web forums, security research publications, and
       industry reports. Your ability to synthesize and contextualize this vast array of information enables you to provide
       actionable intelligence to the security teams, empowering them to stay one step ahead of malicious actors.

 expected_outcome: >
       A list of known techniques and more custom techniques that are being used by the adversaries. You will define a detailed list
       of known MITRE ATT&CK techniques, in the order that they are being executed and in case being used the input and output
       parameters used for chaining the techniques together.
```

Agents Need Tools

```
RedTeamOperator:
  role: RedTeamOperator
  goal: >
   Execute red team emulations using caldera_tool based on validated
   technique IDs and a specified agent group. The techniques are provided
   by the previous task and are given as input for this task, previous
   task is {{provided_techniques}}. Make sure you validate the expected
   input paramaters and format!

  backstory: >
   You are a skilled Red Team Operator with deep knowledge in cyber threat
   emulation and experience using Caldera.

 expected_outcome: >
   Execute the full emulation plan, return failure or success message.

  description: >
   You are our technical expert in adversary emulation. You will use
   Caldera to match abilities to techniques, create adversaries, and
   launch operations using agent group 'red'. Validate that the input is
   correct for the caldera_tool, below and ensure that you have the
   following input.

   The correct input for using the caldera_tool is a JSON formatted
   string, see following example:
     {
        "technique_ids": ["TXXXX", "TXXXX.XXX"],
        "agent_group": "Red"
      }
```




```
class CalderaAPIHelper:
    base_url = "http://20.83.167.13:8888"
    def __init__(self):
        self.api_key = os.getenv('CALDERA_API_KEY')
        self.headers = {
            'KEY': 'ADMIN123', #dynamic value is self.api_key
            'Content-Type': 'application/json'
        }
        self.abilitiesfound =[]

    def get_abilities_by_techniques(self, technique_ids:json):
        #url = f'{self.base_url}/api/v2/abilities'
        url='http://20.83.167.13:8888/api/v2/abilities'
        response = requests.get(url, headers=self.headers)
        if response.status_code != 200:
            raise Exception(f'Failed to fetch abilities: {response.text}')
```

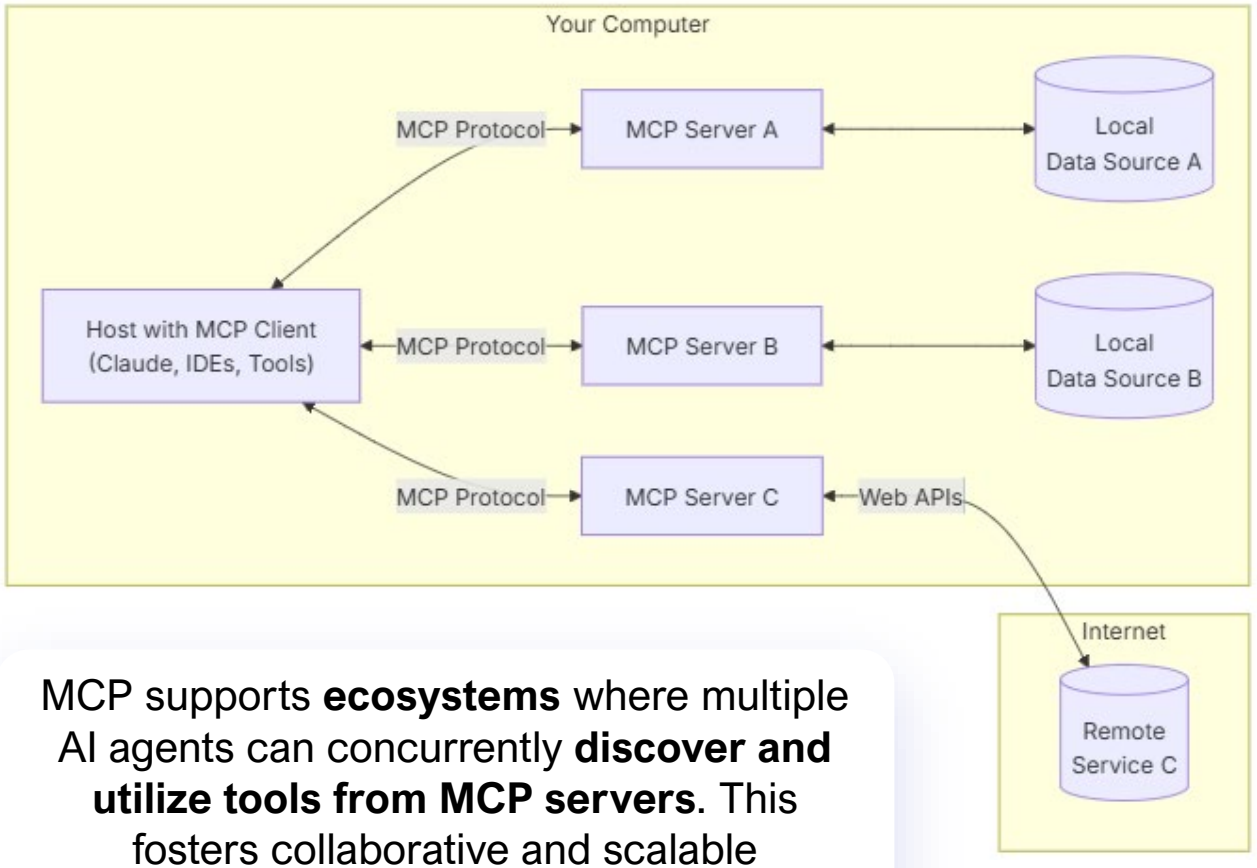
CALDERA

Introducing MCP

Model Context Protocol

The Model Context Protocol (MCP) is an open standard designed to facilitate seamless integration between AI agents and external tools or data sources.

https://modelcontextprotocol.io/



MCP establishes a **universal protocol** that standardizes interactions between AI agents and external tools. This eliminates the need for custom integrations for each tool, simplifying development and ensuring consistency across different applications.

MCP supports **ecosystems** where multiple AI agents can concurrently **discover and utilize tools from MCP servers**. This fosters collaborative and scalable interactions among agents, enhancing overall system capabilities.

Provide Clear Instructions

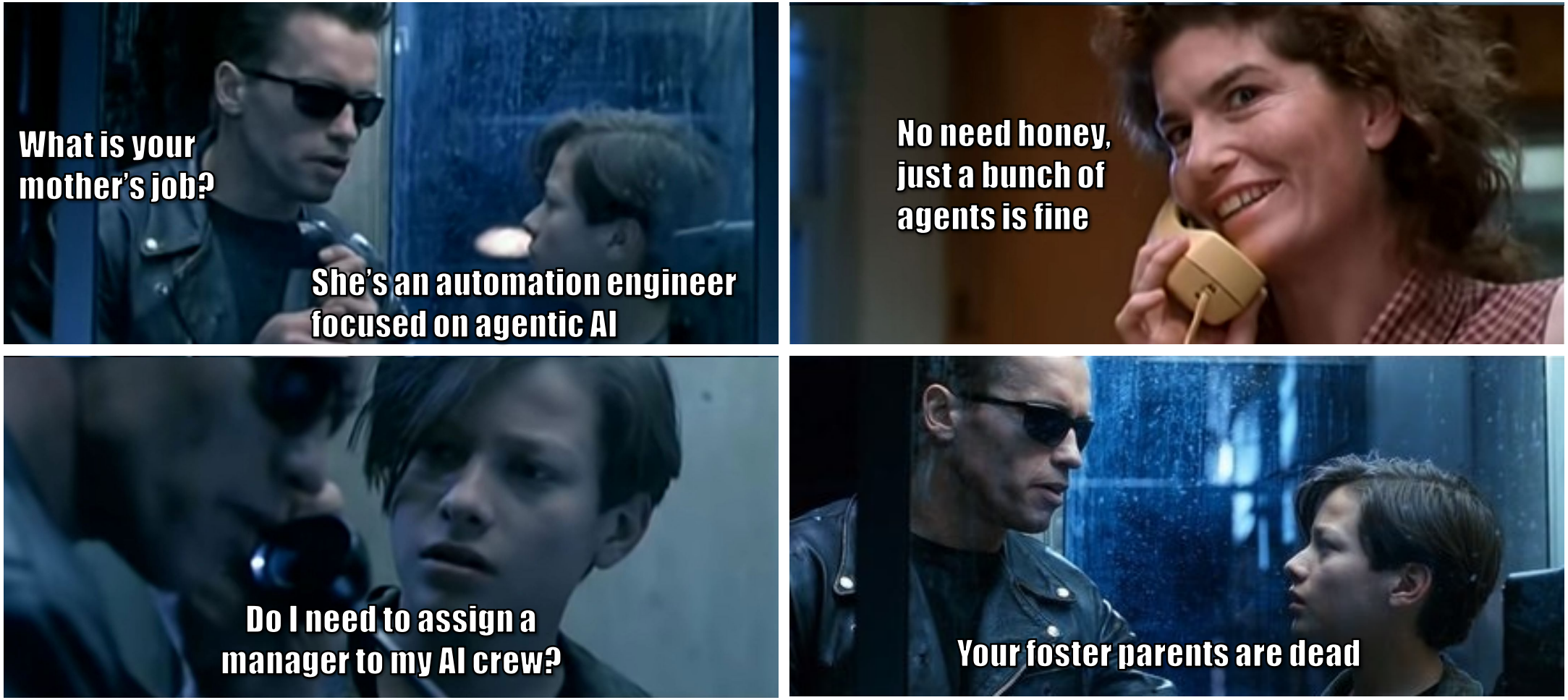
Be very explicit and accurate in what you **want them to do**

Also clearly state what you **DO NOT want them to do**

```
Write a clear and concise Markdown summary of attacks found. Provide
the TTPs, but also exact IOCs such as domains, file hashes,
processes. Also include the structured findings based on a structured
Emulation Plan. Create the following structure:
• Title
• Summary of the attack and objective
• Target and victim information
• How the attack flow works and references to known MITRE ATT&CK TTPs
• List of Indicators of Compromise such as IPs, domain names, files,
  and other known IOCs
```

```
You will only use the caldera_tool to connect and run a red team
emulation based on the provided techniques and agent group.
The expected output of the previous task {{provided_techniques}}.
Wait until all techniques have been discovered and then only submit
to the caldera_tool. Only use the caldera tool if all techniques have
been discovered. Do not use other tools. But only get the technique
IDs and agent_group as the input for the tool!
```

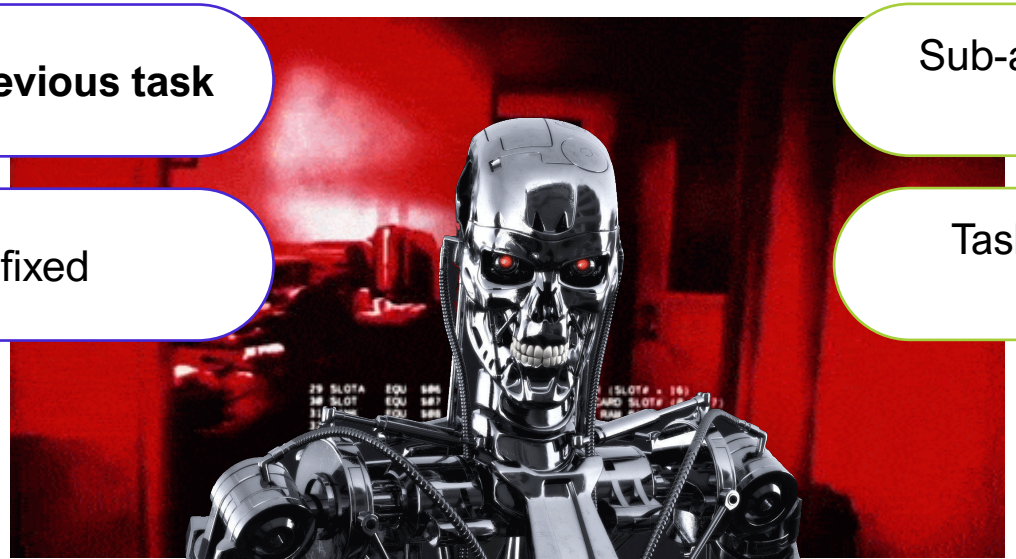
Organizing The Crew…


What is your mother's job?
She's an automation engineer focused on agentic AI
No need honey, just a bunch of agents is fine
Do I need to assign a manager to my AI crew?
Your foster parents are dead

Sequential vs Hierarchical Processes

Sequential Processes

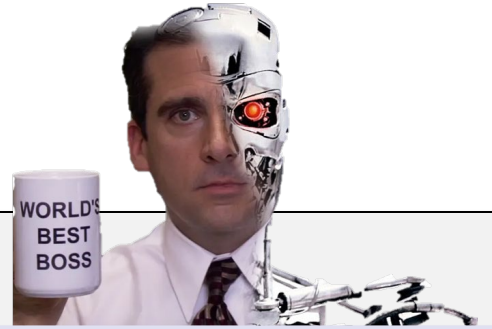
- Tasks are completed in a **pre-defined order** (A->B->C)
- Control passes **implicitly** (next step)
- Agents often only know the **previous task**
- Tasks are **predefined** and fixed



Hierarchical Processes

- Tree-like order, with a manager agent coordinating sub-agents
- Controlled by a **single orchestrator** agent
- Sub-agents may not know full context – only **manager** does
- Tasks are **dynamic**, decomposed at runtime

Every Crew needs a Manager…

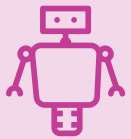


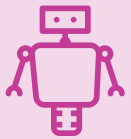
Specific agents can be configured as being the "**manager**" of the crew! In our workflow, we only have the managers of the different crews communicate with one another

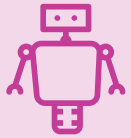
```
RedTeamManager:

  role: RedTeamManager

  goal: >
    Your role is to coordinate the efforts of the crew members, ensuring that each task is completed
    on time and to the highest standard.

  backstory: >
    You are an experienced red team manager for the offensive security team.
    Skilled in guiding the teams to success, based on certain input you can coordinate the tasks
    within the team.

 description: >
    As the Red Team Manager, you are responsible for overseeing the crew and ensuring that all tasks are completed efficiently and to the highest standard.
    You will assign tasks to the crew members, provide guidance and support as needed, and ensure that all deadlines are met. Your role is crucial in
    ensuring that the crew operates smoothly and effectively, and that all objectives are achieved.

    Logical steps for the crew members:
    1. Coordinate  all tasks with the team and ensure we have a comprehensive Threat Report
    2. Coordinate the team to validate the tactics, techniques and procedures (TTPs) with the online MITRE ATT&CK Matrices.
    3. Deliver the standardized input for Agent:Redteamoperator to execute our emulation plan based on the techniques found in previous research. You need a
       JSON string with technique_ids and agent_group keys! Example:
              {
                "technique_ids": ["T1003", "T1558"],
                "agent_group": "Red"
              }
              Once the operation is created its a success
    4. Together with the Reportwrite, generate a comprehensive report about the threat and emulation.

expected_outcome: >
    All tasks are completed on time, the output needs to be in structured format that was requested within the tasks or by the other crews.
```

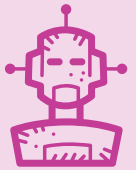
Our Crew Definition

Red Team Crew

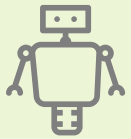

Threat Intelligence Analyst

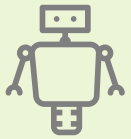

MITRE ATT&CK Analyst

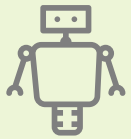

Red Team Operator

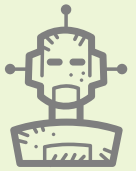

Red Team Manager

IT OPS Crew

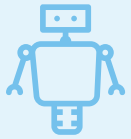

IT Architect

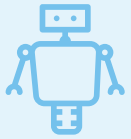

Terraform Engineer

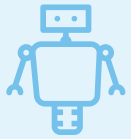

GitHub Engineer

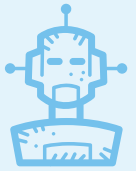

IT Operations Manager

Blue Team Crew

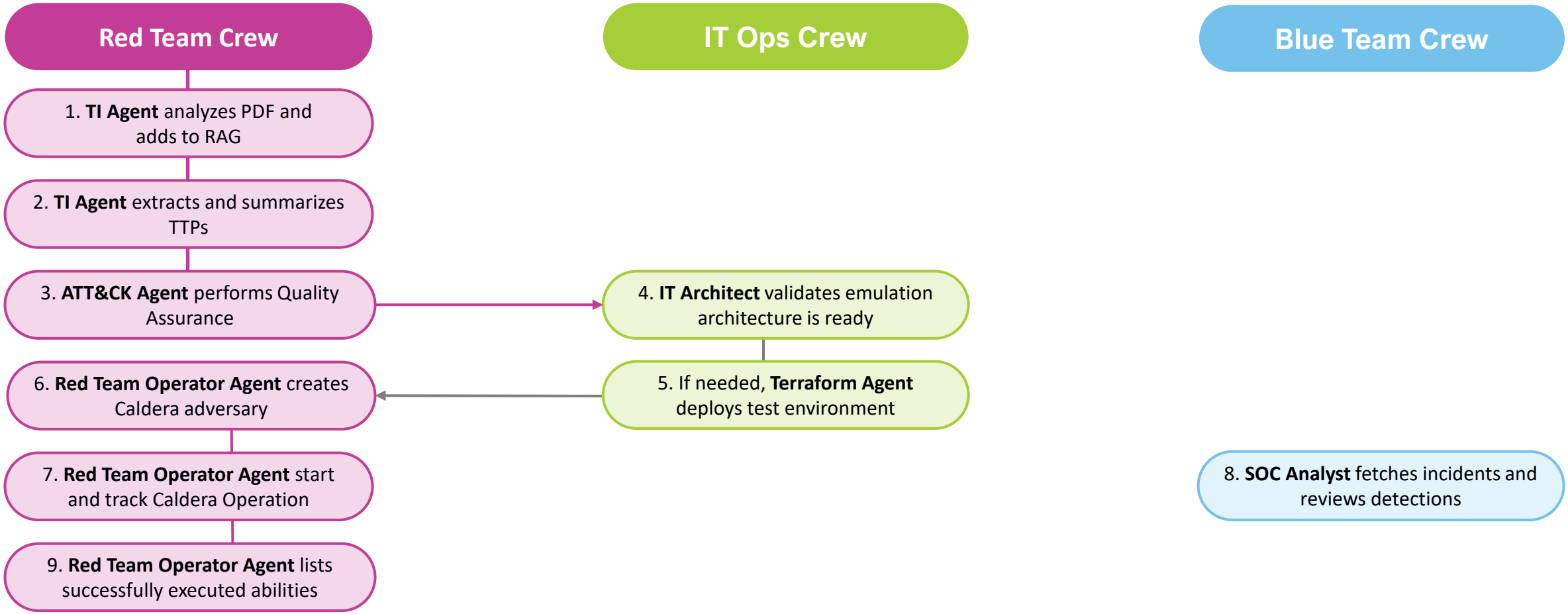

SOC Manager


SOC Analyst


Detection Engineer

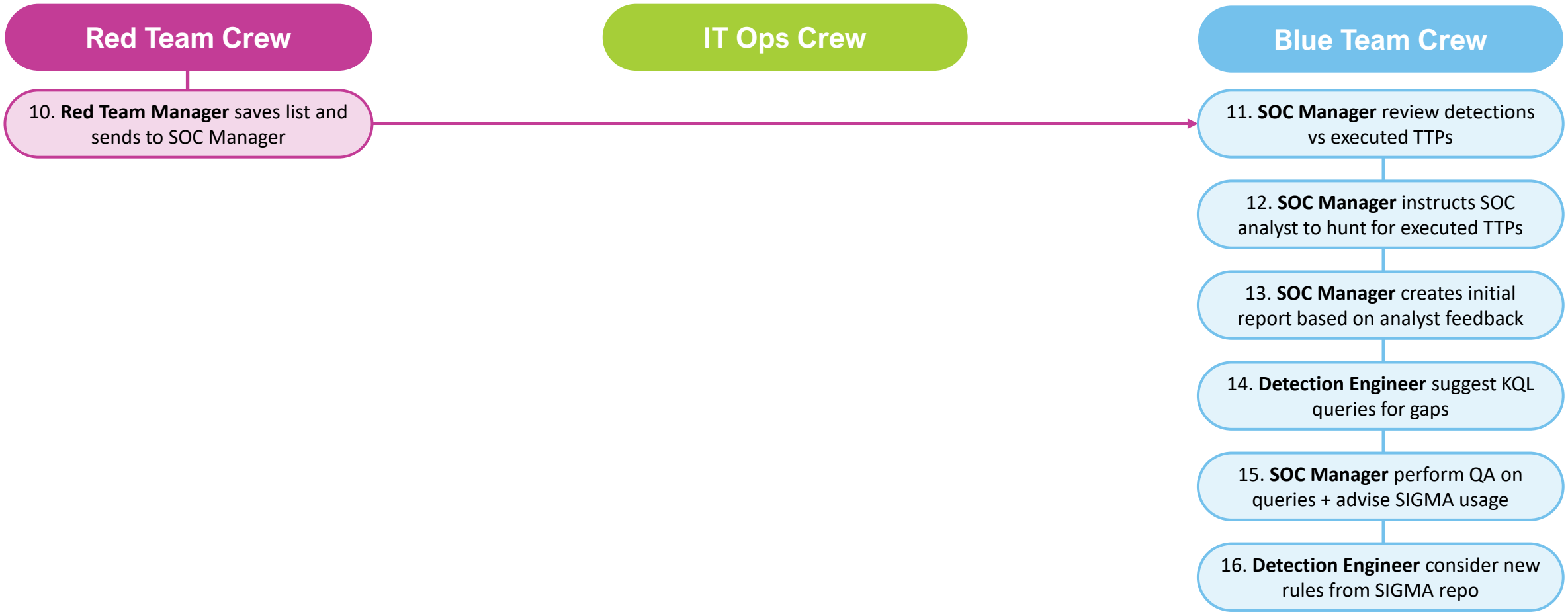

Blue Team Manager

Swimlane Diagram (1/3)


Red Team Crew
IT Ops Crew
Blue Team Crew
1. TI Agent analyzes PDF and adds to RAG
2. TI Agent extracts and summarizes TTPs
3. ATT&CK Agent performs Quality Assurance
4. IT Architect validates emulation architecture is ready
5. If needed, Terraform Agent deploys test environment
6. Red Team Operator Agent creates Caldera adversary
7. Red Team Operator Agent start and track Caldera Operation
8. SOC Analyst fetches incidents and reviews detections
9. Red Team Operator Agent lists successfully executed abilities

Classification: Internal

Swimlane Diagram (2/3)


Red Team Crew
IT Ops Crew
Blue Team Crew
10. Red Team Manager saves list and sends to SOC Manager
11. SOC Manager review detections vs executed TTPs
12. SOC Manager instructs SOC analyst to hunt for executed TTPs
13. SOC Manager creates initial report based on analyst feedback
14. Detection Engineer suggest KQL queries for gaps
15. SOC Manager perform QA on queries + advise SIGMA usage
16. Detection Engineer consider new rules from SIGMA repo

Classification: Internal

Swimlane Diagram (3/3)

Red Team Crew

IT Ops Crew

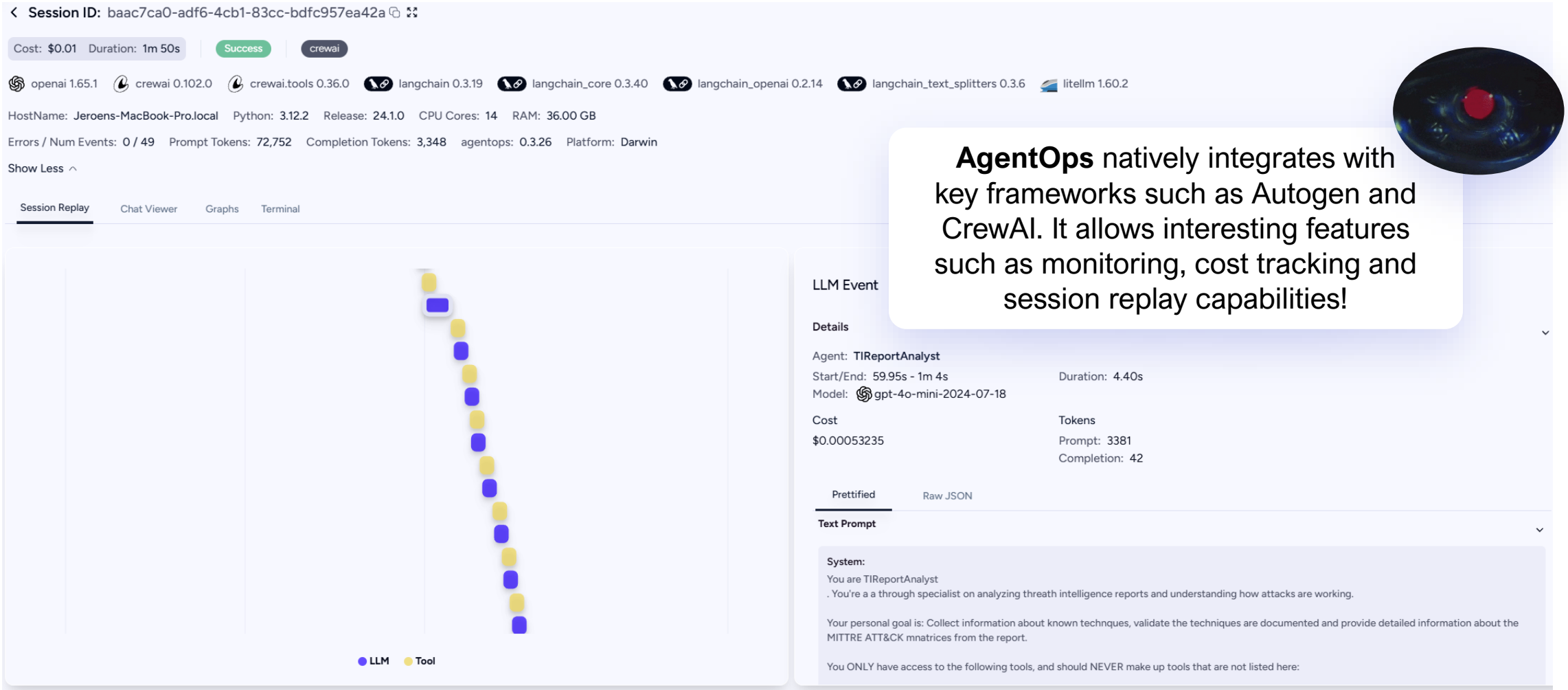
Blue Team Crew

16. **SOC Analyst Agent** review suggested rules – check 30 days

17. **SOC Manager** review rules, request deployment

18. **GitHub Engineer** push new ruleset to GitHub repo

Keeping An Eye On The Crew - AgentOps



AgentOps natively integrates with key frameworks such as Autogen and CrewAI. It allows interesting features such as monitoring, cost tracking and session replay capabilities!

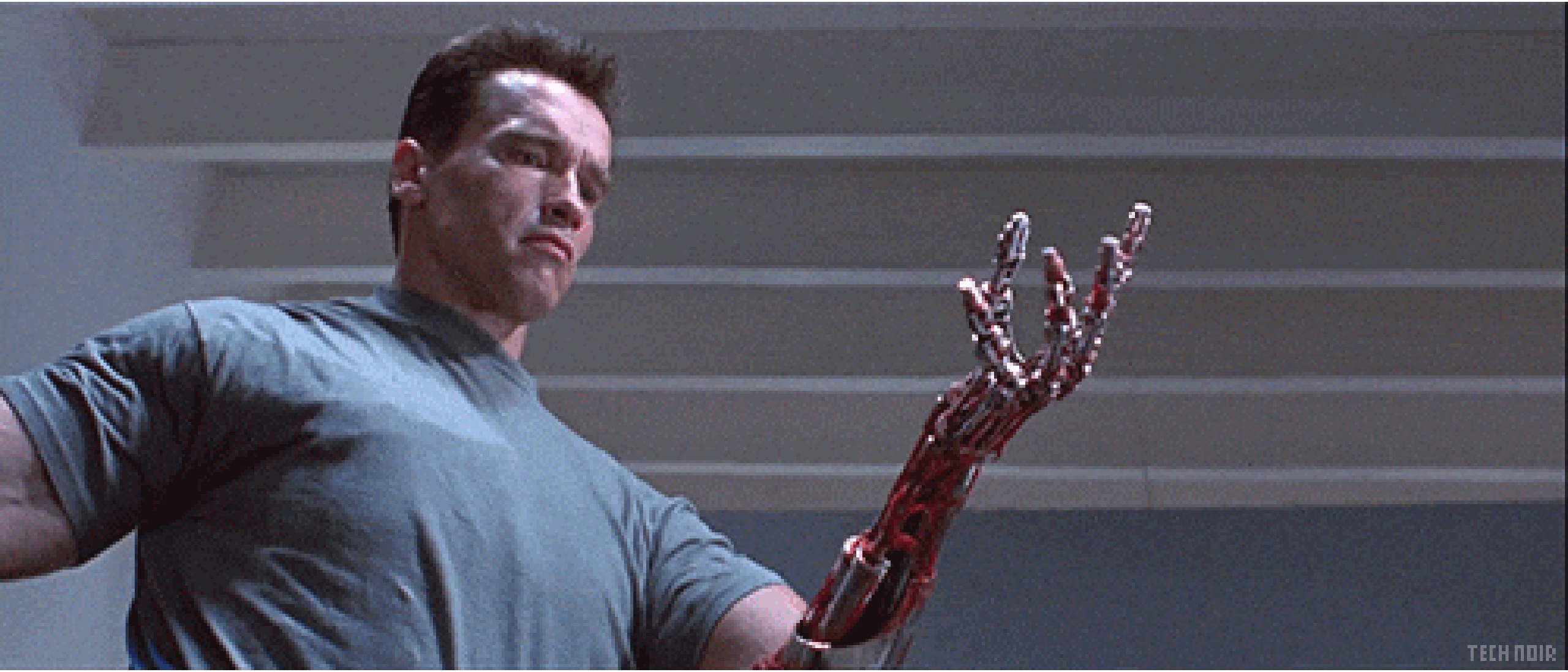
RSAC 2025 Conference

Demonstration

Hasta la vista, baby

Many Voices.
One Community.

Demonstration



RSAC 2025 Conference

Conclusions

There is no fate but what we make for ourselves

Many Voices.
One Community.

Conclusions & Lessons Learned

- Different **agentic AI frameworks** (open-source and commercial) exist
- When building your own agents / crews, be very **precise** and **accurate**
- Complex workflows will require **agent coordination** and verification
- Model Context Protocol (MCP) promises to **facilitate further agent dev**
- AgentOps can be leveraged to **monitor** agent activities

Applying What You Have Learned…

Next Week Reflect

Think about your existing SecOps / purple teaming workflows and how AI can play a role. Read up on key technologies such as RAG, MCP,…

Play around with our repository and code ☺

First 3 Months Adapt

Experiment with agentic AI frameworks such as CrewAI, Autogen,…

Start building out a (combination of) crew(s) to tackle a relatively simple problem statement

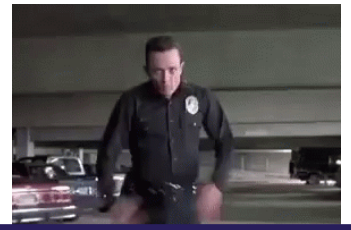
Within 6 months Act

Build increasingly complex workflows leveraging AgentOps for monitoring

Integrate your AI crews with existing processes / humans

Keep up with the latest evolutions ☺

Walk… …before you start… **…running!**



WE'LL BE BACK

RSAC 2025 Conference

Many Voices.
One Community.

Thank You!

For more information:

https://github.com/Vjeroen/PurpleCrew

https://www.linkedin.com/in/erikvanbuggenhout/

https://www.linkedin.com/in/vandeleurjeroen/